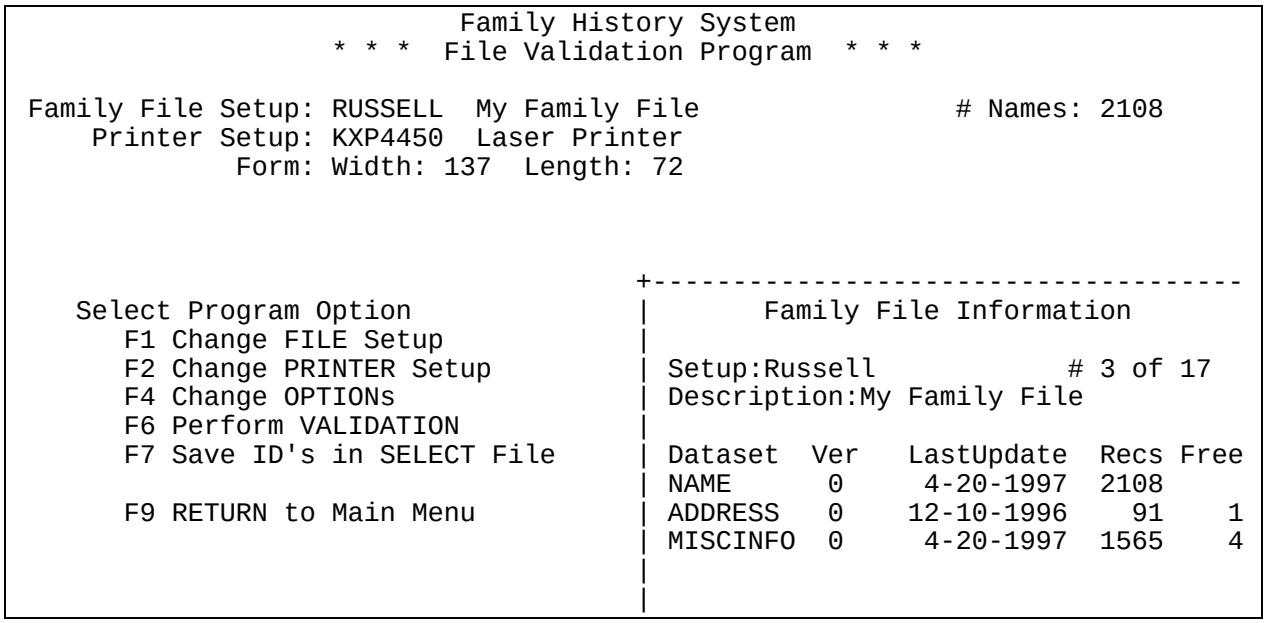
X (F3-E) File Validation

When creating any file of information, errors can be introduced into the file in a variety of ways; sometimes as the result of incorrect information from the source, sometimes because of errors in data entry, and sometimes as a result of program or hardware malfunctions. It is therefore important to periodically examine the information to maintain its accuracy. The purpose of the program that is discussed in this section is to assist you in performing certain types of checks or validations of your family file information. The first option, *Pointer Validation*, is primarily concerned with examining the consistency of various "hidden" data fields that tie the pieces of your family file together. The second option, *Data Validation*, is concerned with checking various fields of your family data for "correctness". This includes checking the consistency of relationships between separate data items (such as birth, death and marriage dates). A third option, *Search for Duplicates*, looks for records, which have the same values in key fields.

All three options are serviced by Main Menu selection F3-E, File Validation. When you choose that option, the screen will be cleared and reformatted as follows:



View 1: File Validation Program

The viewing area in the lower right corner of the screen shows some statistics about the current Family File setup including the numbers of records and "free" records in each of the family file datasets. ("Free" records are records, which have been "deleted" and so are available for reuse when new information is entered.)

As with other programs in the system, option F1 is used to select another Family File Setup and program option F2 is used to select another Printer Setup. Program Option F4 is used to make changes to option settings for each of the validation operations.

When you select program option F6, you will be asked to:

Select 1) Pointer Validation 2) Data Validation 3) Find Duplicates

The operation of each of these options is described in sections X.A, X.B and X.C respectively.

Program option F7 allows you to save, in a SELECT Work file, a list of ID numbers of records for which errors have been found during the most recent validation procedure. This file can then be used by other programs to perform other operations on those records. For instance, you could use it in the Search/Select/LIST program to produce a simple list of information about the error ID's, in the Family Group report program to print a set of Family Group reports for the error ID's, or in the file update program (Main Menu option F1-F4-F3) to review each record and make corrections.

X.A (F6-1) POINTER VALIDATION

To permit you to place variable amounts of "miscellaneous" information and comments about an individual in your family files without requiring excessive amounts of space on your diskette or hard disk, an individual's information is stored in many "records" distributed among 3 datasets (name, address and miscellaneous information). These might be thought of as 3 "card files" storing different types of information. All of these records are "drawn together" by a collection of system maintained *pointer* fields. You are no doubt familiar with some of these fields. The mother and father ID numbers in the name record are two of them, and the spouse ID in a marriage record is another. There are many others that you are not (and need not be) aware of. In addition, each record in the family file datasets has an identifying code for the type of information stored in the record and, for "subordinate" records, the "source" of the information (whether an address record is an individual or family residence, for example). It is important that the complete set of "pointers" and record identifiers for those records relating to an individual be valid and consistent.

The first record of each family file dataset also has information (the DATE and TIME that the file was originally CREATEd) that helps the programs determine whether the datasets that you are using "belong together". This is to protect you from inadvertently attempting to enter information into, or produce reports using datasets that are from two or more family files that you may be working with at different times. In the ".ADR" and ".OTH" datasets, these "header records" also have some additional "hidden pointers" that help the file update program keep track of deleted records.

There are several ways in which inconsistencies may be introduced into this "hidden" collection of information. Because updated file records may remain in memory and not be written to disk until the files are closed (or until you return to the Main Menu program from the file update program) any interruption of an update session by prematurely turning off your computer or by a power failure may result in incomplete updates to the system information. Errors in some versions of the file update program are another (unfortunate) source of inconsistencies in system pointers (for instance, at one time you could enter the same ID number for mother and father, which would introduce errors in the files). The purpose of the procedure invoked by Program Option "F6-1" is to detect and eliminate any errors in the system maintained "hidden" information.

NOTE: *If a family file has been corrupted as a result of an interrupted update session and a recent backup is available, it is always preferable to restore the file from the backup, since correction using the validation program may result in some loss of information. If there have been many updates to the file since the last backup, or the file errors are also in the backup copy, then the validation program can provide a "clean" file for you to continue your work with minimal loss of data.*

When you select the "Pointer Validation" option, the viewing area in the lower right corner of the screen will be cleared and reformatted as shown at right. This area will be used to show the numbers of errors of each type that are found while the validation operation is in progress.

```
Pointer Validation Error Summary
     0  Name Record Errors
     0  Spouse Record Errors
     0  Address Record Errors
     0  Comment Record Errors
     0  Event Record Errors
     0  Miscellaneous Record Errors
     0  Duplicate References
     0  Unreferenced Records
     0  Other Errors
     0  TOTAL Errors
```

View 2: Pointer Validation Error Summary

At this time you will be asked whether to check all records in the file or just SELECTed records. If you choose to check selected records then you will be prompted to enter the name of a previously created SELECT Work file. **NOTE:** *If you do not choose to process the full file, then some error conditions will not be checked... in particular there will be no attempt to determine whether there are "unreferenced" records in the .ADR and .OTH datasets.*

You will also be asked if you want to:

Make Corrections? (Y/N)

The error conditions that will be checked and reported are the same whether corrections are made or not. The types of corrections that are made by the program are intended to eliminate the error condition, usually by setting an incorrect pointer to zero. This does not recover any information or reestablish any relationships that may have been affected by the bad pointer. You will want to examine, using the file update program (Main Menu selection F1) any records reported as having errors to determine whether any information should be reentered or relationships reestablished.

Finally, you will be asked to:

Select 1) Screen 2) Printer 3) File

to identify the destination for the error report. If you choose "Screen" then the errors will only be printed as single line messages at the bottom of the screen. The program will pause after each error message is encountered. You can continue the program by pressing the space bar (or any other character key on the keyboard). If you continue the processing by pressing the PGDN key, then the program will not pause for other error messages. If you choose "Printer" or "File" as the report destination, then a formatted report will be sent there and the error messages will be simply "echoed" on the bottom line of the screen, without pausing. In either case, the count of errors of different types will be shown in the viewing area in the lower right corner of the screen.

After determining the destination for the error messages, the validation process begins. Actually this validation procedure is divided into several phases. More will be said about what is done in each phase later but they may be described briefly as follows:

- Validate information in Name records (this is by far the most time consuming of the phases, taking perhaps 90% of the total execution time); this is the only phase of checking performed if you do not perform the validation for the full file
- Check for broken "sibling" chains (determine children who are not "listed" as a child of a recorded parent)
- Check Address dataset free record chain
- Check for isolated or unreferenced address records
- Check Miscellaneous dataset free record chain
- Check for isolated or unreferenced miscellaneous records
- Synchronize date time stamps in "header" records if files are "unmatched" and updates are being performed.

There is one type of "error" detection and correction that is controlled by a procedure option which is set using program option F4-1. That is the reporting of individuals (name records) who are unusually aged, based upon their birth date and the fact that their date of death is zeroes. These are likely to be persons whose date of death is unknown or unrecorded. The Family History System provides a method of dealing with such conditions, so that they are not reported as living to an unrealistically advanced age. If an individual's year of death is "9999" then the report programs will interpret that as a "don't know" value... that is they will assume that the person has died but the date of death is unknown. The age of these individuals will not be calculated. Among the option settings for the Pointer Validation procedure are two which control

whether or not the check is performed and what is the maximum age that would be accepted without assuming an individual had died. (The default setting for these options in the distributed report definition file is "N"o Checking and 125 years.) Even if the error checking is turned on, the dates of death will not be modified unless you have allowed the program to "Make Corrections".

For those who are interested in customizing the pointer validation report or who would like a list of the different error messages that the report may contain, use Main Menu option F3-B-F5-F2 and select report PVLD. Then use F6 to print a report of the parameters that are used to produce the pointer validation report. The error messages are defined as Report Variables with variable names: EMxx. The report variables with names of the form: RTxx are words describing the various types of records in the FHS family file. These words are inserted into some of the error messages that may apply to more than one type of record.

The suggested procedure for using this program is:

- Backup your family file datasets. If none of the datasets exceed the capacity of the diskettes you use, the DOS COPY command (or Windows File Manager) can be used to back them up, otherwise you will have to use the DOS (Windows) BACKUP command (or some equivalent utility) to back them up
- Run option "F6-1" of this program with output going to the printer or a file and allowing the program to make file changes
- Use program option F7 to create a SELECT work file of ID numbers that have had an error reported for them
- Run option "F6-1" again with output going to the printer (or a different report file)... all *program correctable* errors should now be gone
- Use the file update program (Main Menu option F1-F4-F3) and the validation reports to review information for individuals whose ID# appeared in any of the error messages.

The next section gives more detailed descriptions of the various phases of the validation process. This information is primarily for the few users who may be interested in this level of detail concerning the pointer validation process. You do *not* have to read or understand any part of that discussion to make use of this option.

X.A.1. Pointer Validation Details

Before discussing the phases of the pointer validation process in detail, let's first look at the type of "identity" information stored in the "prefix" of each family file record. Each record begins with a 1 character record type as shown in the "RTYPE" table on the next page. The record types were changed in the December 1985 update of the system. Prior to that the record types were numeric or alphabetic characters. At that time they were changed to decimal codes which do not correspond to standard characters. This program recognizes both the old and new record types, but changes all old record types to the new type when updating the datasets. In addition, the "header" record of each family file dataset has a leading 1 character file type which is used to check that the dataset has been properly initialized. This "file type" is also shown in the table below.

	Old RTYPE		New RTYPE
	CHAR	ASCII	ASCII
Name record	1	049	001
Address record	2	050	002
Spouse Record	3	051	003
Place Record	4	052	004
Comments	F	070	005
Education record	7	055	007
Work record	8	056	008
Military record	9	057	009
Medical record	A	065	010
NAME dataset	N	078	078 (unchanged)
ADDRESS dataset	A	065	065 (unchanged)
MISC dataset	M	077	077 (unchanged)

Table 1: FHS Record & File Identification Codes

In addition, each address and misc info record contains the record type and record number of the "source" record to which it is appended. In the error messages produced by this program, the "source record type" is labeled "SRTYPE"

and the "source record number" is labeled "SRNO". (For example a "family" residence will have SRTYPE=3, the record type of a spouse record. A comment record attached to an Education record will have SRTYPE=7.)

Next we will look at the detailed processing that takes place during each phase of Pointer Validation.

I. During "Phase I" of the validation process, each name record is read successively and the following checks are performed:
   1. Mother and Father ID
      a. must be between 0 and the highest ID# on record (invalid parent ID's are set = 0)
   2. Sibling chain
      a. the name record for the oldest child is retrieved and parent ID's checked to make sure that ID# of name record being validated is either the father ID (FID) or mother ID (MID)
      b. younger children's records are retrieved and parent ID's are similarly verified (if the ID# of the name record being validated is not found as a parent in a child's record, the sibling chain is terminated)
      c. if a child's ID is encountered a second time while following the sibling chain, the "loop" is noted and the sibling chain is terminated
      d. a note is made of each child correctly located on sibling chain; this information will be used later for identifying "broken" sibling chains.
   3. Birth/Death Place Record
      a. record number must be between 0 and max Misc Rec #
      b. if record number>0 then record is retrieved and record type and source record information is checked (see previous table of record types and discussion of source record information)
   4. Comment Records for individual
      a. first comment record ID must be between 0 and max Misc rec#
      b. if comment records are present, first comment record is retrieved and record type and source record information checked; total comment record count from first record is saved; backward pointer should be 0
      c. successive comment records are retrieved and record prefix verified as for first record; backward pointer should point to previous record
      d. after last comment record is retrieved, total record count is compared to what had been stored in first comment record (if record type or source record information is incorrect, the comment chain is terminated; all other discrepancies are corrected by the program)
   5. Address Records for individual
      a. first address record # must be between 0 and max addrs rec#
      b. if address records are present, first address record is retrieved and record type and source record information checked
      c. successive address records are retrieved and record prefix verified as for first record; (if record type is incorrect it is corrected; if source record information is incorrect, the address chain is terminated)
      d. address comment records are checked (as in 3.)
   6. Spouse information
      a. first spouse record # must be between 0 and max misc rec#
      b. if spouse records are present; each spouse record is retrieved and checked for valid record type and source record information
      c. spouse ID's in the record are checked to see if one corresponds to the name record being validated; (if record type or source record information is incorrect, or ID# is not in spouse record, the spouse record chain is terminated)
      d. if marriage place record is present for any of them, that record is retrieved and record type and source record information is checked

e. spouse comment records are checked (as in 3.)

f. spouse residence records are checked (as in 4.)

7. Miscellaneous information

a. first misc record # of each type must be between 0 and max misc rec#

b. if misc information is present, first record is retrieved and record type and source record information is checked; (if record type or source record information is incorrect the chain is terminated for that type of misc info);

c. misc info comment records are validated (as in 3.)

d. misc info address information is validated (as in 4.)

II. After all name records have been individually checked, the record of all validated parent pointers is checked to see if any name record was not located on a parent-child chain. (Unverified parent ID's are set =0 in the name record)

III. Records on free chain of Address dataset are checked to see if they have been referenced during Phase "I". Address records which were unreferenced in Phase "I" and not on FREE record chain are noted and added to FREE record chain. Count of FREE records in the address dataset header record is compared to the number of records on the FREE chain. A discrepancy is noted and corrected

IV. Records on free chain of Miscellaneous Info dataset are checked to see if they have been referenced during Phase "I". Records in Misc Info dataset which were unreferenced in Phase "I" and not on FREE record chain are noted and added to FREE record chain. Count of FREE records in the misc info dataset header record is compared to the number of records on the FREE chain. A discrepancy is noted and corrected.

While checking for unreferenced misc records, the number of references (during Phase "I") to each spouse record is checked. A message is displayed if a spouse record has both spouse ID's nonzero but was not referenced exactly 2 times. You must then use the file update program (main menu option F1) to retrieve the marriage record from the spouse ID from which it IS accessible, delete the marriage record, and then re-add it.

X.B (F6-2) DATA VALIDATION

Program option F6-2, the *Data Validation* option, is used to look for incorrect or unreasonable information in a family file. Actually, this option might be more aptly described as a "DATE Validation" option, because most of the error checking involves the many dates that can be entered into an FHS family file. But the checking extends beyond merely determining whether a date represents a valid calendar date. In addition, the program attempts to determine whether two separate dates have a proper "logical" relationship to one another. For instance, a birth and death date in a NAME record should represent a reasonable "age span" for the individual represented by the record, and the birth dates of a child's parents should bear a "reasonable" relationship to the birth date of the child.

X.B.1. DATA VALIDATION OPTIONS

Program option F4-2 allows you to change settings that control many of the types of error checking that will take place during the Data Validation process. When you select this option, the lower right portion of the screen is cleared and reformatted with a list of parameters for the Data Validation Error Report and procedure. The table at the top of the next page lists the options and their default settings. To change the setting for an option, use the UP/DOWN cursor control keys to move the "reverse video" hiliting to the setting that you wish to change, press the Enter key, type the desired setting value and press Enter again. You may save the option settings in the Report Definition File by pressing the F1 key to end the option update process. If you end the process by pressing the ESCape key, then any changes to the option settings will only be temporary.

Default	----- Validation Options -----
Y	Check Name Record Dates (Y/N)
Y	Check Marriage Dates (Y/N)
Y	Check Address Dates (Y/N)
Y	Check Other Record Dates (Y/N)

Y	Check for Invalid Age (Y/N)
Y	Check Spouse Age (Y/N)
60	Max Spouse Age Difference
Y	Check Age at Marriage (Y/N)
15	Min Age at Marriage
90	Max Age at Marriage
Y	Check Age at Child Birth (Y/N)
14	Min Age at Birth of Child
60	Max Age at Birth of Child
Y	Check SEX Codes (Y/N)
Y	Check SEX Code of Parent (Y/N)
Y	Check SEX Code of Spouse (Y/N)
Y	Check for Blank Marriage (Y/N)
Y	Check for Unmarried Parents
Y	Check Father-Child Surnames
Y	Use Soundex for Surname Check

Table 2: Data Validation Options

The types of error checking that can take place are:

- *Check for Invalid Dates* - determine that the mm, dd and yyyy parts of a date represent a true calendar date and that the date is not greater than the "current date". All types of dates may be checked, including birth and death dates, beginning and ending marriage dates, event dates, beginning and ending address dates, and beginning and ending dates in each of the "miscellaneous" record types: Medical, Educational, Military and Occupational. Invalid dates also include death dates which precede birth dates, and ending dates which precede beginning dates in marriage, address, or miscellaneous record types
- *Check for Unreasonable Ages* - the program computes a person's age at certain key events (death, marriage, etc.) and compares this to values that you have identified as representing a reasonable age at which the event could occur. There are four types of age checking:
  - ⇒ *Age of each Individual* - determine that the birth and death dates (or birth and current date) yield a "reasonable" value for the person's age. **NOTE:** The Pointer Validation option can eliminate many of the very large ages that appear in reports by setting a missing death date to "00-00-9999"
  - ⇒ *Age at Marriage* - determine whether a person's age at marriage was either "unreasonably" young or old
  - ⇒ *Ages of Spouses* - determine whether there is an "unreasonable difference" between the ages of two married persons
  - ⇒ *Ages of Parents* - determine whether a parent was unreasonably young or old at the birth of a child.
- *Check for Unmarried Parents* - determine whether each parent is married to the other parent of each child; this is not an "error" condition, but may be useful information when planning to export information using a GEDCOM file because parent-child relationships are defined under the marriage or family record of the parents
- *Check for Blank Marriage Records* - look for marriage records which have only a single spouse recorded, no dates or places, and no addresses or comments...these are records which were probably created accidentally by starting to add or create a marriage record, deciding against it, and then terminating the process by pressing the F1 key (thus SAVEing the empty record) rather than using the ESCape key
- *Father-Child Surnames* - determine whether a child has either the same surname as the father, or one which "sounds similar" to the father's; this is a common custom in the U.S. so variances may indicate that a parent-child relationship has been improperly defined.
- *Gender Codes* - valid gender codes are taken from the GENDER System Table. Gender code checks include:
  - ⇒ *Gender Code of Individual* - determine whether each Name record contains a valid gender code, as indicated in the GENDER System Table

⇒ *Gender of Parent* - determine whether each father is a Male and each mother is a Female; The program will not "correct" an "error condition" of this type because it cannot determine which is incorrect: the Gender code of the parent or the Parent ID in the child's record. (Also, this may not be an "error" condition at all, according to your records, if one of the parents were an adoptive parent)

⇒ *Gender of Spouse* - report marriages in which both spouses are of the same sex. This is not an "error" condition in an FHS family file, but could cause problems during export to a GEDCOM file because the GEDCOM specification requires identifying participants using gender specific labels of Husband and Wife.

In the descriptions of the types of error checking that will take place, there were repeated references to "reasonable" or "unreasonable" values for various ages. Of course you may have been concerned about just what determined an "unreasonable" value. To avoid making "unreasonable" assumptions about "reasonable" values, you are permitted to enter your own definitions of "reasonableness" as values for several of the validation options.

You may choose to include any combination of these types of error checking during a validation process. The reason for not performing a type of error checking may be to simply concentrate on one or a few types of errors, or to avoid the overhead of checking certain types of errors when you know that no errors of those types exist.

X.B.2. Performing the DATA VALIDATION Procedure

The Data Validation process is begun by selecting program option F6-2. You will first be asked to:

Select Validation for 1) Full File 2) SELECTed records

If you press "2" then you will be prompted to:

Enter SELECT File Dataset Name: SELECT.WRK

This SELECT file must have been previously created; for example, by the Search/Select/LIST program or perhaps by the F7 option of the Validation program. A line will be shown indicating the number of records that will be checked and the total number of records in the family file.

You will next be asked if you want to:

Select: 1) Screen 2) Printer 3) FILE

as the destination for the error report. If you select "1" then the errors will only be shown as one line messages on the bottom line of the screen, otherwise they will be sent as part of a "Data Validation Report" to the chosen destination. If you select "3", then you will also be asked to:

Enter REPORT File Name: REPORT .FIL

to identify the dataset that will receive the report. The dataset will be placed on the drive and in the directory of the Report Group in the File Name Table (see Section VI)

Even when the error report destination is a printer or file, the error messages will be echoed on the bottom line of the screen. However the program will not pause after each error message as it would if the report destination were the screen.

The first stage in a data validation process (unless you are only checking surnames) is to build tables of birth years and sex codes for all name records in the family file. This stage is noted by the message:

Building Birth Year Table at hh:mm:ss

at the bottom of the screen. The progress of the table creation, is shown by a counter in the lower right corner of the screen. The purpose of the tables is to reduce the number of times that a name record may have to be retrieved when checking the relationships between various data items.

NOTE: *The ages at time of marriage, the ages of parents at time of birth of a child, and the age differences of spouses are actually determined from the birth years of the individuals involved. The month and day of birth are not used because we are only interested in "reasonable" values, not the exact values of ages.*

During the validation process, the ID number of the record being checked is shown in reverse video in the lower right corner of the screen. As "errors" are found, running totals of the numbers of each type of error are shown in the table of error types in the lower right portion of the screen. A message describing each error is shown on the bottom line of

the screen, preceded by a reverse video display of the Name Record ID number to which the message applies. If the report destination is the Screen, then the program will pause until a key is pressed. If the ESCape key is pressed, the validation process will be terminated. If the PGDN key is pressed, the validation process will be placed in NOPause mode during which processing is continued *without pausing* after each subsequent message. If the program is in NOPause mode, you can cause it to pause by pressing any key on the keyboard. If you continue the program again by pressing a key other than ESCape or PGDN, then the program will again pause after each error message is displayed.

X.B.3. DATA VALIDATION REPORT

The Data Validation Report is produced when the destination for "errors" is a printer or a file. The report is described by the DVLD Report entry in the Report Definition File. You can use Main Menu option F3-B-F5-F2 to list the reports and select the DVLD report definition. The report consists of several parts:

- *Report Preface* - which lists each type of data validation that will be performed, based upon the option settings. This portion of the report is defined by HTF (Heading/Title/Footing) lines in the report definition. There is a separate "Title" line for each type of data validation
- *Error Messages* - which describe the individual errors found by the program. The ID number of the Name Record for which the error message was produced is placed at the beginning of the line (and will *not* be repeated on successive error messages for the same ID). The error messages are constructed from report variables with names of the form "EMxx". These "Error Message" variables include references to other report variables that will be replaced by specific values when the message is printed in the report
- *Report Summary* - which gives the total number of errors of each type that were found during the data validation process. The Total Lines are constructed from report variables with names of the form "ETxx".

Although this is called an "Error Report", it is really only intended to call attention to conditions which may indicate that a date or relationship is incorrect. You should use the file update program, Main Menu option F1, to examine each ID number that appears in the report to determine whether there is an error and to make corrections if necessary. You can save the list of ID numbers for which errors were encountered using program option F7. This list can be retrieved by the File Update program by using Main Menu option F1-F4-F3-Enter. Individual Name records can then be selected for display/update from the list of ID numbers.

X.C (F6-3) FINDING DUPLICATE RECORDS (an Extended Option for Registered Users)

The procedure described in this section was requested by a user who was faced with the task of trying to identify possible duplicate records resulting from importing a GEDCOM file into an existing family file. What he wanted to do was compare each new record to each old record to determine whether they were the same (or nearly the same) in certain key fields such as name, gender, birth/death dates and places. After examining the problem, it was apparent that the task would be greatly simplified if the records were already grouped together according to the values of those key fields...that is, if they were placed in sorted sequence according to the values of those fields.

One of the extended options of the Family History System allows processing family records in a sorted sequence by using "Index Files". (See Chapter XI) An Index File is just a list of ID's that have been placed in a sorted sequence. However, the "sort fields" for which index files could be created were rather limited. The sort fields did not include birth & death places, and the options for sorting dates were not very flexible. To support the procedure for identifying possible duplicate records, the procedure for creating Index Files (Main Menu selection F3-F of the extended system) was enhanced to allow:

- additional sort fields of Birth & Death Place
- sorting of text fields by actual value or by soundex code (the use of "soundex codes" could result in "similar sounding" names being grouped together)

- options for choosing only portions of a date for a sort field (for example. YYYY or YYYYMM) to allow dates which are "almost the same" to be placed together

The steps for performing a search for duplicate records will then include the following:

- Identify the two groups of records that are to be compared (for example, compare records 1-2345 to records 2346-2579).

  NOTE: *Either or both of the groups could be the full file*

- Create an Index File which uses the fields for which duplicate values are to be found as sequence fields for the Index File
- Use Validation Option F3-E-F6-3 to look for duplicate records.

You may want to follow this with program option F7 to Save the list of "duplicate" ID's to a SELECT work file and then use the Search/Select/LIST option to produce an indexed detail list of the "matching" records.

During the process of looking for "matching" name records, a message will appear on the bottom line of the screen describing each group of matching records. You may also request that a report be created summarizing the results of the search procedure. That report can be sent to a printer or to a report file. If you have used the Soundex value for text fields for determining matches, then this Duplicate Entries Report will show both the Soundex code and a "representative value" for those text fields. The "representative values" are taken from the first of the records in the group and should not be assumed to be the actual value for all records in the group.

X.C.1. Performing the DUPLICATE DATA Search

The procedure for finding records which have the same or "similar" values for certain key data items is begun by selecting program option F6-3 of the File Validation Program. Before choosing this option you should have created an Index File (using Main Menu selection F3-F of the extended system) for which you have selected the fields that will be searched for "matching" values as sort fields. You will also have identified at that time whether the text fields will be sorted on all or part of the actual text field or on the soundex value of the text field, and whether date fields will be sorted on all or part of the date field (month/day/year or combinations). Your decisions there will have been recorded in the header record for the Index File and will be used during the "Duplicate Date Search" in this program.

When you choose program option F6-3 the lower right corner of the screen will be cleared and reformatted as shown at right. The "Compare Fields" section will be used to show the data items for which "matching values" will be searched. The "Comparison Group Size" will show the number of records in the two groups of records that will be compared and the "Minimum Matched Group Size" will show the minimum number of "matching records" that must be in a group for it to be reported. The "Total" line will show a summary of results while the search is being performed.

```
        Duplicate Data Summary

Compare Fields:
 * Surname   * Given Name  _ Sex Code
Date:  * Birth  _ Marriage  _ Death
Place: _ Birth  _ Death
Comparison Group Size:
   Group #1: _1234  Group #2: __345
Minimum Matched Group Size: _2

Total Matches: __________Groups: ____
```

View 3: Duplicate Data Summary

You will also be prompted to:

Enter Matched Field INDEX File Name: _____

where the name of the index file will default to that of the Index File that you have just created. After pressing the Enter key, the index file will be opened, the header record will be read and the "Compare Fields" portion of the Duplicate Data Summary will be filled out with an * for each data item that was used as a sort field.

You will next be asked to identify the two groups of records that will be compared when looking for "matching data". First you will be prompted to choose:

Group #1: 1) Full File 2) Relation File 3) Select File 4) ID Range

If you choose "Full File" then the "Comparison Group Size" for Group #1 will be filled in with the total number of name records in the family file. If you choose "Relation File" or "Select File" then you will be prompted to enter the dataset name for a previously created relationship work file (ANCESTOR/ DESCNDNT/ RELATIVE.WRK) or SELECT work file and the number of ID's in those work files will be shown as the size of Group #1. If you choose "ID Range" for Group #1, you will be prompted to:

Enter Range of ID's: First = ____1 Last = _2108

where the default range will encompass the full family file. If you are comparing original family records to newly imported family records, then the "Last" number should be changed to the highest ID number that existed in the file prior to the import operation. When you press the enter key, the size of Group #1 in the "Duplicate Data Summary" will be filled in with the number of ID's in the chosen range of ID's.

You will next be prompted in a similar way for the records that will be in comparison group #2. If you choose "ID Range" again for the comparison group, the First/Last ID's will default to a range of ID's that complements the one chosen for Group #1.

Finally, you will be asked to identify the destination for the Duplicate Data summary report. If you choose "Screen" as the destination, then the report will only consist of lines that will be printed at the bottom of the screen for each group of matching records and the "Total" information that is shown in the "Duplicate Data Summary" in the lower right portion of the screen. The program will pause after each group of "marching" records is found. You may then terminate the search process by pressing the ESCape key, allow the search process to continue without further pause by pressing the PGDN key, or continue the search process with a pause after the next matching group of records is found by pressing any other key.

If you choose "Printer" or "File" for the report destination, then the information for groups of "matching" records will be shown on the screen as described above except that the program will not pause after a group of matching records is reported.

X.D (F7) SAVING THE ERROR IDs IN A SELECT FILE

Program option F7 allows you to create a SELECT work file containing a list of the ID's for which "errors" were found during a validation procedure. This can be used in the Search/Select/LIST option or in the Family Group Report options to print a report of information about the individuals for whom errors were found. Or you can use it in the file update program (Main Menu selection F1-F4-F3) to assist you in examining the records for which "errors" were reported. You may also use the SELECT file to perform a data validation again, perhaps after you have made some corrections, and avoid having to reprocess all the records that were found to be error free during the first validation process.

NOTE: *If the validation procedure that preceded the creation of the Select File was the procedure for finding duplicate records, then the records in the two groups are identified as primary and secondary selections respectively, which would allow you to list them separately in reports produced by the Search/Select/LIST option.*